Biostat 200C Homework 3

Due May 19 @ 11:59PM

Zian ZHUANG

Q1.

The **log-logistic** distribution with the probability density function

$$f(y) = \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2}$$

is sometimes used for modelling survival times.

- (a) Find the survivor function $S(y)$, the hazard function $h(y)$ and the cumulative hazard function $H(y)$.

Answer:

Firstly we can get $F(y)$,

$$\begin{aligned} F(y) &= \int_0^y f(y)dy \\ &= \int_0^y \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2}dy \\ &= |_0^y \frac{e^{\theta}y^{\lambda}}{1+e^{\theta}y^{\lambda}} \\ &= \frac{e^{\theta}y^{\lambda}}{1+e^{\theta}y^{\lambda}} \end{aligned}$$

Then it is easy to get $S(y)$,

$$\begin{aligned} S(y) &= 1 - F(y) \\ &= 1 - \frac{e^{\theta}y^{\lambda}}{1+e^{\theta}y^{\lambda}} \\ &= \frac{1}{1+e^{\theta}y^{\lambda}} \end{aligned}$$

Then we can have $h(y)$,

$$\begin{aligned} h(y) &= \frac{f(y)}{S(y)} \\ &= \frac{e^{\theta}\lambda y^{\lambda-1}}{(1+e^{\theta}y^{\lambda})^2} / \frac{1}{1+e^{\theta}y^{\lambda}} \\ &= \frac{e^{\theta}\lambda y^{\lambda-1}}{1+e^{\theta}y^{\lambda}} \end{aligned}$$

Then we can calculate $H(y)$,

$$\begin{aligned} H(y) &= -\log S(y) \\ &= -log(\frac{1}{1+e^{\theta}y^{\lambda}}) \\ &= log(1+e^{\theta}y^{\lambda}) \end{aligned}$$

- (b) Show that the median survival time is $\exp(-\theta/\lambda)$.

Answer: We know that the median survival time means $F(y) = 0.5$, then we can get,

$$\begin{aligned} 0.5 = F(y) &= \frac{e^{\theta}y^{\lambda}}{1+e^{\theta}y^{\lambda}} \\ \implies e^{\theta}y^{\lambda} &= 1 \\ y &= \exp(-\theta/\lambda) \end{aligned}$$

- (c) Plot the hazard function for $\lambda = 1$ and $\lambda = 1$ with $\theta = -5$, $\theta = -2$, and $\theta = 1/2$, in one figure.

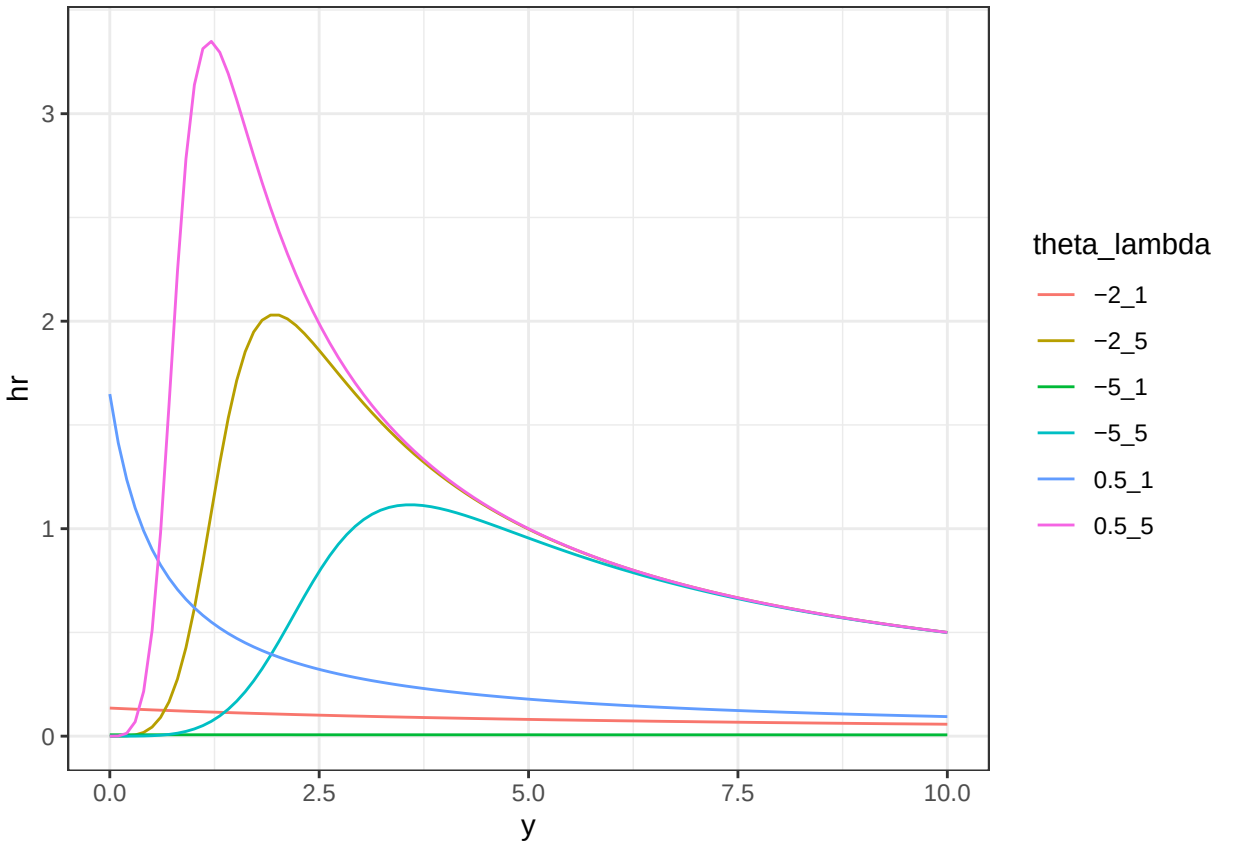
Answer:

```
.hf <- function(y, theta, lambda){
  hr <- (exp(theta)*lambda*y^(lambda-1))/(1+exp(theta)*y^(lambda))
  return(hr)
}

plot_data <- NULL

for(theta in c(-5,-2,0.5)){
  for(lambda in c(1,5)){
    plot_data <- rbind(plot_data,
                       data.frame(hr = .hf(y=seq(0,10,length=100),theta,lambda),
                              y=seq(0,10,length=100),
                              theta_lambda = paste0(theta,"_",lambda)))
  }
}

ggplot(data=plot_data, aes(x=y, y=hr, group=theta_lambda, color=theta_lambda)) +
  geom_line() + theme_bw()
```



Q2. ELMR Exercise 7.5

The data arise from a large postal survey on the psychology of debt. The frequency of credit card use `ccarduse` is a three-level factor ranging from never, occasionally to regularly.

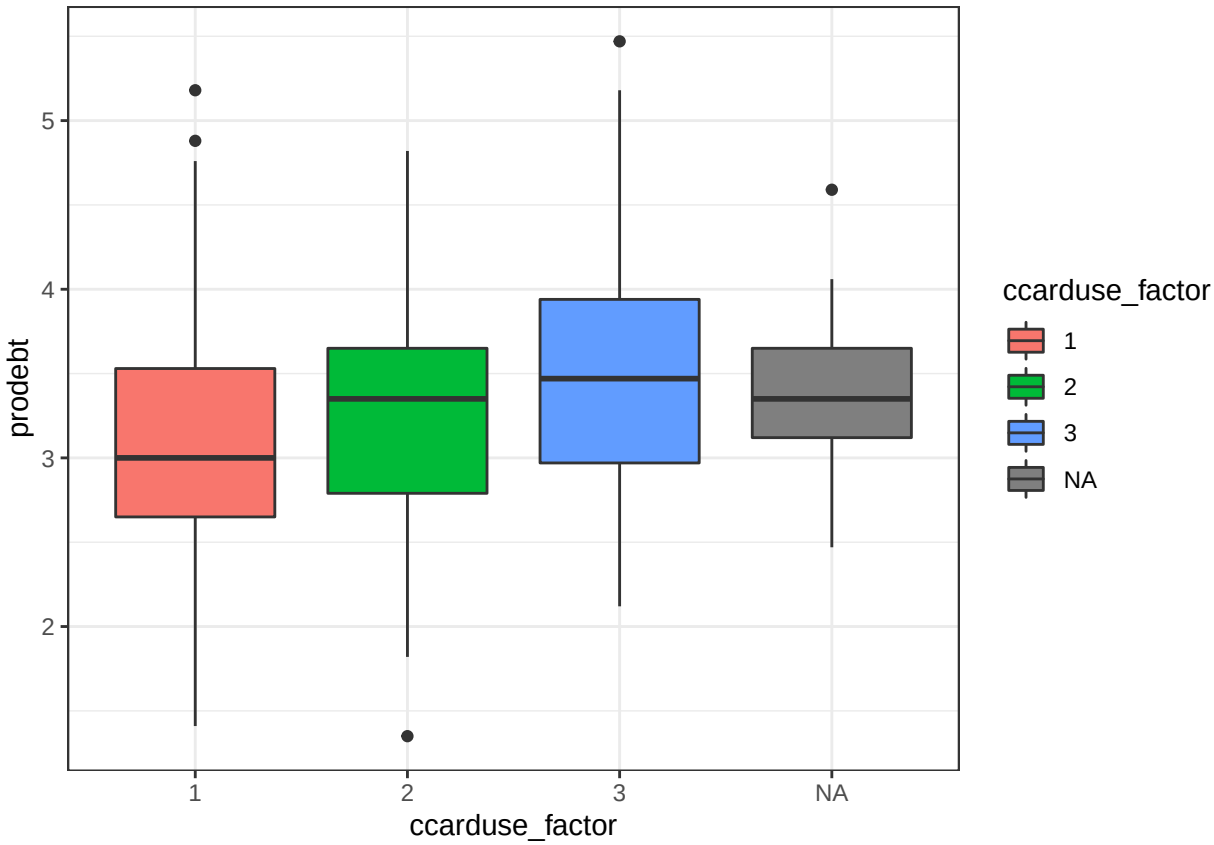
```
data(debt)
#help(debt)
```

- (a) Declare the response as an ordered factor and make a plot showing the relationship to `prodebt`. Comment on the plot. Use a table or plot to display the relationship between the response and the income group.

```
df <- debt %>%
  mutate(ccarduse_factor = as.factor(ccarduse)) %>%
  mutate_at(vars(incomegp,house,agegp),ordered) %>%
  select(-ccarduse)

ggplot(df, aes(x=ccarduse_factor, y=prodebt, fill=ccarduse_factor)) +
  geom_boxplot() + theme_bw()
```

```
## Warning: Removed 45 rows containing non-finite values (stat_boxplot).
```

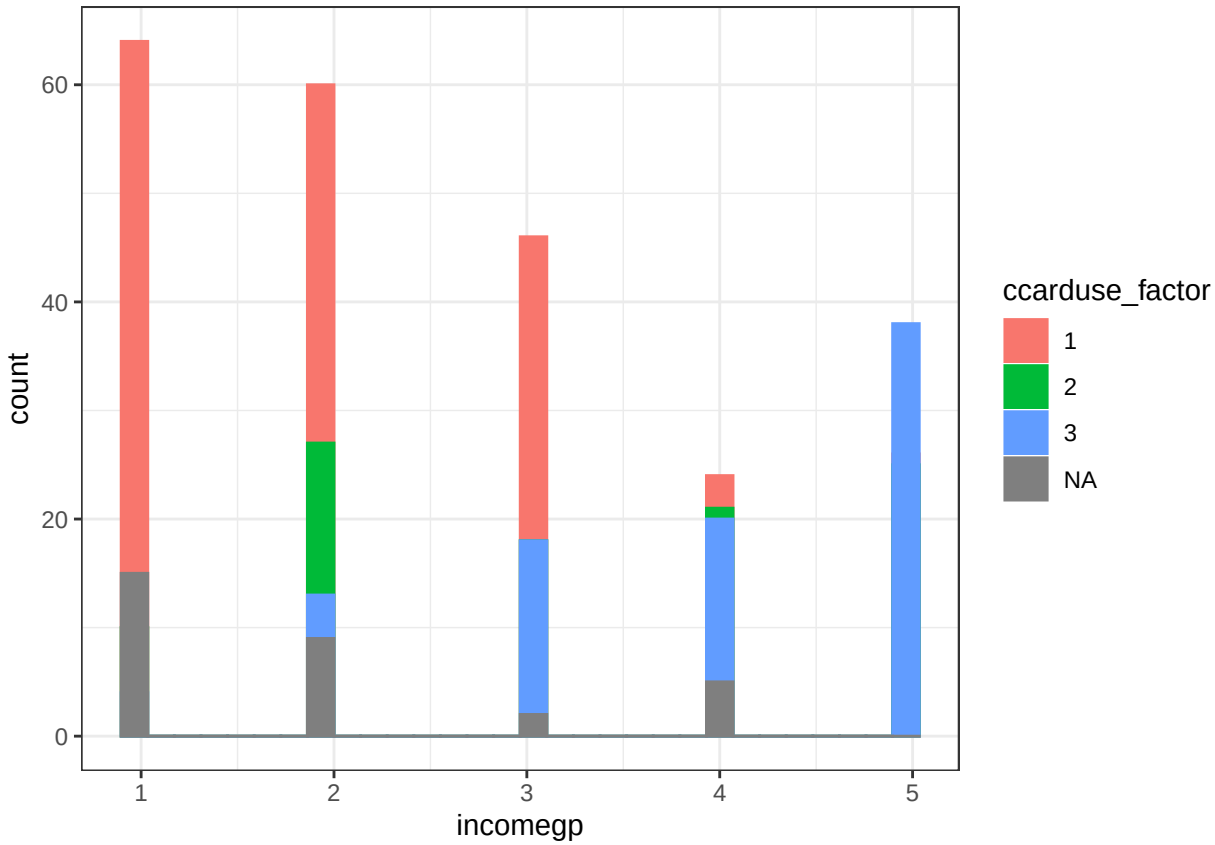


As we can tell from the box plot, a higher frequency of a person use credit cards are associated with a higher attitude to debt.

```
# the relationship between the response and the income group
ggplot(df %>% mutate_at(vars(incomegp),unclass) %>%
         mutate_at(vars(ccarduse_factor),as.factor),
       aes(x=incomegp, color=ccarduse_factor, fill=ccarduse_factor)) +
  geom_histogram(position = 'identity') + theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 19 rows containing non-finite values (stat_bin).
```



As we can tell from the box plot, a higher frequency of a person use credit cards are associated with a higher level of income.

- (b) Fit a proportional odds model for credit card use with all the other variables as predictors. What are the two most significant predictors and what is their qualitative effect on the response? What is the least significant predictor?

```
debt$ccarduse <- as.factor(debt$ccarduse)

pofit <- polr(ccarduse ~., data = debt)
summary(pofit)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = ccarduse ~ ., data = debt)
##
## Coefficients:
##              Value Std. Error t value
## incomegp  0.47131     0.1061  4.4423
## house     0.11600     0.2324  0.4992
## children -0.07872     0.1250 -0.6296
## singpar   0.88172     0.5971  1.4766
## agegp     0.20568     0.1576  1.3050
```

```
## bankacc   2.10270     0.5934  3.5435
## bsocacc   0.47322     0.2671  1.7715
## manage    0.18179     0.1653  1.0998
## cigbuy   -0.73546     0.2981 -2.4674
## xmasbuy   0.47014     0.4130  1.1385
## locintrn  0.11881     0.1424  0.8344
## prodebt   0.61046     0.1822  3.3497
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  7.9694  1.4752     5.4023
## 2|3  9.3944  1.5051     6.2417
##
## Residual Deviance: 511.673
## AIC: 539.673
## (160 observations deleted due to missingness)
```

```
## store table
(ctable <- coef(summary(pofit)))
```

```
##
## Re-fitting to get Hessian
```

```
##                Value Std. Error    t value
## incomegp  0.47131302  0.1060967  4.4422968
## house     0.11600148  0.2323630  0.4992251
## children -0.07872411  0.1250325 -0.6296291
## singpar   0.88171828  0.5971140  1.4766330
## agegp     0.20568368  0.1576103  1.3050145
## bankacc   2.10269577  0.5933918  3.5435203
## bsocacc   0.47321630  0.2671328  1.7714643
## manage    0.18179169  0.1652902  1.0998331
## cigbuy   -0.73545858  0.2980681 -2.4674178
## xmasbuy   0.47014289  0.4129631  1.1384622
## locintrn  0.11881236  0.1423979  0.8343685
## prodebt   0.61046374  0.1822466  3.3496579
## 1|2       7.96937466  1.4751711  5.4023391
## 2|3       9.39436162  1.5051079  6.2416532
```

```
## calculate and store p values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

## combined table
(ctable <- cbind(ctable, "p value" = round(p,3)))
```

```
##                Value Std. Error    t value p value
## incomegp  0.47131302  0.1060967  4.4422968   0.000
## house     0.11600148  0.2323630  0.4992251   0.618
## children -0.07872411  0.1250325 -0.6296291   0.529
## singpar   0.88171828  0.5971140  1.4766330   0.140
## agegp     0.20568368  0.1576103  1.3050145   0.192
## bankacc   2.10269577  0.5933918  3.5435203   0.000
```

```
## bsocacc   0.47321630  0.2671328  1.7714643   0.076
## manage    0.18179169  0.1652902  1.0998331   0.271
## cigbuy   -0.73545858  0.2980681 -2.4674178   0.014
## xmasbuy   0.47014289  0.4129631  1.1384622   0.255
## locintrn  0.11881236  0.1423979  0.8343685   0.404
## prodebt   0.61046374  0.1822466  3.3496579   0.001
## 1|2       7.96937466  1.4751711  5.4023391   0.000
## 2|3       9.39436162  1.5051079  6.2416532   0.000
```

As we can tell from the results that `incomegp` and `bankacc` are most significant predictors. The least significant predictor is `house`.

```
exp(ctable[,1]) %>% round(.,3)
```

```
##  incomegp     house  children   singpar     agegp   bankacc   bsocacc    manage
##     1.602     1.123     0.924     2.415     1.228     8.188     1.605     1.199
##    cigbuy   xmasbuy  locintrn   prodebt       1|2       2|3
##     0.479     1.600     1.126     1.841  2891.049 12020.414
```

Interpret: Increase level in income group is associated with a higher odds of using credit cards more often. People have a bank account are associated with a higher odds of using credit cards more often.

- (c) Fit a proportional odds model using only the least significant predictor from the previous model. What is the significance of this predictor in this small model? Are the conclusions regarding this predictor contradictory for the two models?

```
pofit_s <- polr(ccarduse ~ house, data = debt, Hess=TRUE)
summary(pofit_s)
```

```
## Call:
## polr(formula = ccarduse ~ house, data = debt, Hess = TRUE)
##
## Coefficients:
##       Value Std. Error t value
## house 0.558     0.1433   3.895
##
## Intercepts:
##     Value  Std. Error t value
## 1|2 1.3000 0.3118     4.1698
## 2|3 2.4670 0.3274     7.5344
##
## Residual Deviance: 847.2661
## AIC: 853.2661
## (36 observations deleted due to missingness)
```

```
## store table
(ctable <- coef(summary(pofit_s)))
```

```
##           Value Std. Error  t value
## house 0.5579957  0.1432559 3.895097
## 1|2   1.2999505  0.3117572 4.169753
## 2|3   2.4669924  0.3274285 7.534447
```

```r
## calculate and store p values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

## combined table
(ctable <- cbind(ctable, "p value" = round(p,3)))
```

```
##           Value Std. Error  t value p value
## house 0.5579957  0.1432559 3.895097       0
## 1|2   1.2999505  0.3117572 4.169753       0
## 2|3   2.4669924  0.3274285 7.534447       0
```

```r
exp(ctable[,1]) %>% round(.,3)
```

```
##  house    1|2    2|3
##  1.747  3.669 11.787
```

In this small model, individuals with mortgage house (house=2) or owned outright (house=3) are significant ($p<0.001$). Nevertheless, it not necessarily contradicts the previous model, since effect of house might be explained by other predictors in the the previous model.

- (d) Use stepwise AIC to select a smaller model than the full set of predictors. You will need to handle the missing values carefully. Report on the qualitative effect of the predictors in your chosen model. Can we conclude that the predictors that were dropped from the model have no relation to the response?

```r
#exclude all na first
df_1 <- df %>% na.exclude
pofit <- polr(ccarduse_factor ~., data = df_1, Hess=TRUE)
pofit_step <- step(pofit)
```

```
## Start:  AIC=539.41
## ccarduse_factor ~ incomegp + house + children + singpar + agegp +
##     bankacc + bsocacc + manage + cigbuy + xmasbuy + locintrn +
##     prodebt
##
##            Df    AIC
## - house     2 537.66
## - manage    1 538.42
## - locintrn  1 538.62
## - xmasbuy   1 538.69
## - singpar   1 538.99
## <none>        539.41
## - children  1 539.65
## - bsocacc   1 540.24
## - agegp     3 543.62
## - cigbuy    1 543.93
## - incomegp  4 544.41
## - bankacc   1 550.29
## - prodebt   1 552.08
##
## Step:  AIC=537.66
```

```
## ccarduse_factor ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + locintrn + prodebt
##
##            Df    AIC
## - locintrn  1 536.54
## - manage    1 536.90
## - xmasbuy   1 537.25
## - singpar   1 537.38
## <none>        537.66
## - children  1 538.33
## - bsocacc   1 539.59
## - cigbuy    1 542.31
## - agegp     3 542.56
## - incomegp  4 545.98
## - bankacc   1 550.69
## - prodebt   1 550.99
##
## Step:  AIC=536.54
## ccarduse_factor ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + manage + cigbuy + xmasbuy + prodebt
##
##            Df    AIC
## - manage    1 535.98
## - singpar   1 536.23
## - xmasbuy   1 536.35
## <none>        536.54
## - children  1 537.30
## - bsocacc   1 538.74
## - agegp     3 541.04
## - cigbuy    1 541.15
## - incomegp  4 545.82
## - prodebt   1 549.07
## - bankacc   1 550.68
##
## Step:  AIC=535.98
## ccarduse_factor ~ incomegp + children + singpar + agegp + bankacc +
##     bsocacc + cigbuy + xmasbuy + prodebt
##
##            Df    AIC
## - singpar   1 535.26
## <none>        535.98
## - xmasbuy   1 536.01
## - children  1 537.00
## - bsocacc   1 539.37
## - agegp     3 540.98
## - cigbuy    1 541.40
## - incomegp  4 544.24
## - prodebt   1 547.33
## - bankacc   1 551.37
##
## Step:  AIC=535.26
## ccarduse_factor ~ incomegp + children + agegp + bankacc + bsocacc +
##     cigbuy + xmasbuy + prodebt
##
```

```
##            Df    AIC
## <none>        535.26
## - xmasbuy   1 535.40
## - children  1 536.07
## - bsocacc   1 538.38
## - cigbuy    1 540.85
## - agegp     3 541.03
## - incomegp  4 542.24
## - prodebt   1 547.06
## - bankacc   1 549.99
```

```
## store table
(ctable <- coef(summary(pofit_step)))
```

```
##                   Value Std. Error    t value
## incomegp.L  1.36178736  0.3752783  3.6287404
## incomegp.Q -0.30366786  0.3296987 -0.9210466
## incomegp.C  0.17033446  0.2979528  0.5716826
## incomegp^4 -0.11255364  0.2669090 -0.4216930
## children   -0.22971420  0.1384326 -1.6593940
## agegp.L     0.54780154  0.3534472  1.5498821
## agegp.Q    -0.91670483  0.3033751 -3.0216878
## agegp.C     0.03771217  0.2393286  0.1575748
## bankacc     2.05036439  0.5839770  3.5110359
## bsocacc     0.59831786  0.2660694  2.2487284
## cigbuy     -0.82847286  0.3060577 -2.7069172
## xmasbuy     0.59537826  0.4132318  1.4407852
## prodebt     0.65625341  0.1804001  3.6377663
## 1|2         4.74921603  0.9752994  4.8694957
## 2|3         6.20183021  1.0026405  6.1854972
```

```
## calculate and store p values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

## combined table
(ctable <- cbind(ctable, "p value" = round(p,3)))
```

```
##                   Value Std. Error    t value p value
## incomegp.L  1.36178736  0.3752783  3.6287404   0.000
## incomegp.Q -0.30366786  0.3296987 -0.9210466   0.357
## incomegp.C  0.17033446  0.2979528  0.5716826   0.568
## incomegp^4 -0.11255364  0.2669090 -0.4216930   0.673
## children   -0.22971420  0.1384326 -1.6593940   0.097
## agegp.L     0.54780154  0.3534472  1.5498821   0.121
## agegp.Q    -0.91670483  0.3033751 -3.0216878   0.003
## agegp.C     0.03771217  0.2393286  0.1575748   0.875
## bankacc     2.05036439  0.5839770  3.5110359   0.000
## bsocacc     0.59831786  0.2660694  2.2487284   0.025
## cigbuy     -0.82847286  0.3060577 -2.7069172   0.007
## xmasbuy     0.59537826  0.4132318  1.4407852   0.150
## prodebt     0.65625341  0.1804001  3.6377663   0.000
## 1|2         4.74921603  0.9752994  4.8694957   0.000
## 2|3         6.20183021  1.0026405  6.1854972   0.000
```

```
exp(ctable[,1]) %>% round(.,3)
```

```
## incomegp.L incomegp.Q incomegp.C incomegp^4   children    agegp.L    agegp.Q
##      3.903      0.738      1.186      0.894      0.795      1.729      0.400
##    agegp.C    bankacc    bsocacc     cigbuy    xmasbuy    prodebt        1|2
##      1.038      7.771      1.819      0.437      1.814      1.928    115.494
##        2|3
##    493.652
```

Interpret:

- Increase level in income group is associated with a higher odds of raise one level of credit cards use.
- Increase number of children in household is associated with a lower odds of raise one level of credit cards use.
- Increase level in age group is associated with a higher odds of raise one level of credit cards use.
- Compared to people have no bank account, people have a bank account are associated with a higher odds of raise one level of credit cards use.
- Compared to people have no building society account, people have a building society account are associated with a higher odds of raise one level of credit cards use.
- Compared to people do not buy cigarettes, people who buy cigarettes are associated with a lower odds of raise one level of credit cards use.
- Compared to people do not buy Christmas presents for children, people who buy Christmas presents for children are associated with a higher odds of raise one level of credit cards use.
- Increase level in attitudes to debt is associated with a higher odds of raise one level of credit cards use.
- (e) Compute the median values of the predictors in your selected model. At these median values, compare the predicted outcome probabilities for both smokers and nonsmokers.

```
df_2 <- data.frame(i=0)
for(i in c("incomegp", "children", "agegp", "bankacc",
           "bsocacc", "cigbuy","xmasbuy", "prodebt")){
  temp <- data.frame(i=quantile(df_1[,i],0.5,type=c(3)))
  df_2 <- cbind(df_2, temp)}
df_2 <- df_2[,-1]
names(df_2) <- c("incomegp", "children", "agegp", "bankacc",
                 "bsocacc", "cigbuy","xmasbuy", "prodebt")
# non smoker
df_2[6] <- 0
predict(pofit_step, df_2, type = "probs")
```

```
##         1         2         3
## 0.3111442 0.3476305 0.3412253
```

```
# smoker
df_2[6] <- 1
predict(pofit_step, df_2, type = "probs")
```

```
##         1         2         3
## 0.5084236 0.3071005 0.1844760
```

As we can tell from the results that the smokers are associated with a lower frequency of credit cards use (never), compared to that of non smokers.

- (f) Fit a proportional hazards model to the same set of predictors and recompute the two sets of probabilities from the previous question. Does it make a difference to use this type of model?

```
df_ph <- df_1 %>%
  dplyr::select(c("incomegp", "children", "agegp", "bankacc",
                  "bsocacc", "cigbuy","xmasbuy", "prodebt", "ccarduse_factor"))
phmodel = polr(ccarduse_factor  ~ ., method = "cloglog", data = df_ph)
summary(phmodel)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = ccarduse_factor ~ ., data = df_ph, method = "cloglog")
##
## Coefficients:
##               Value Std. Error t value
## incomegp.L  0.68409    0.19851  3.4461
## incomegp.Q -0.06668    0.17171 -0.3883
## incomegp.C  0.13240    0.17039  0.7770
## incomegp^4 -0.05261    0.15764 -0.3337
## children   -0.15850    0.08048 -1.9694
## agegp.L     0.34171    0.19949  1.7129
## agegp.Q    -0.53383    0.16983 -3.1433
## agegp.C     0.01662    0.14784  0.1124
## bankacc     1.00245    0.25444  3.9399
## bsocacc     0.36193    0.16242  2.2284
## cigbuy     -0.40915    0.16315 -2.5078
## xmasbuy     0.14582    0.23781  0.6132
## prodebt     0.42334    0.11563  3.6612
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  2.1228  0.5391     3.9377
## 2|3  2.9816  0.5469     5.4513
##
## Residual Deviance: 515.1315
## AIC: 545.1315
```

```
# non smoker
df_2[6] <- 0
predict(phmodel, df_2, type = "probs")
```

```
##         1         2         3
## 0.3692502 0.2937532 0.3369966
```

```r
# smoker
df_2[6] <- 1
predict(phmodel, df_2, type = "probs")
```

```
##         1         2         3
## 0.5003375 0.3052067 0.1944558
```

Proportional hazards model provides a largely consistent results, but with a higher AIC value, which shows that model fits the data worse.